**Project Proposal**

**Introduction**

This document presents the first version of our project proposal for *the 2016 Projects on Big Data Software* class. It outlines scope and objectives of the project and gives an overview of project team, used technologies, datasets and provisional project schedule.

**Project working title**

Parallelization and NoSQL Benchmarking

**Project description**

The project team plans to build a distributed *Extract Transform Load (ETL)* tool using Spark that will process large data input files and populate data into NoSQL Open Source databases such as Cassandra or MongoDB. The team will implement and test performance statistics in the context of the following environments:

* Single node
* multiple nodes (2-4)

The project team will also use Cassandra and MongoDB as targets and benchmark how these NoSQL databases perform under the same load conditions. Various read/write queries will be used to conduct studies on both Cassandra and MongoDB.

**Project objectives**

Main objective: To conduct a benchmark study of parallelization for different NoSQL databases with special emphasis on the following aspects:

* Distributed ETL and how parallelization helps and performs under various conditions
* Performance Testing and Benchmarking of various Open Source NoSQL databases

**Project Team**

* Srinivasa Gorijavolu (Project Lead), IU Username: *sgorijav*, IU Github: *sgorijav*
* Michael Mzyk, IU Username: mmzyk, IU GitHub: mmzyk
* Satish Patharkar, IU Username: SPathark, IU Github: SPathark

**List of Technologies**

* Spark
* Cassandra
* MongoDB

**Compute Resources**

OpenStack in FutureSystems

**System Requirements**

* Size: 3 VM instances
* Storage: 6 GB

**List of DataSets**

Population from the college data set or Dummy Insurance data

**Development language**

Python will be used wherever necessary as development language.

Spark and NoSQL databases (mongodb and/or Cassandra) used to demonstrate parallelism.

**Schedule**

* March 7th : Initial Meeting
* March 12th : Project proposal creation
* March 15th : Project proposal finalization
* March 16th : Project proposal submission
* March 19th : First run
* March 23rd : Build systems
* March 26th : Develop modules, test run
* March 27th : Second run
* March 28th - 30th : Collaboration of results
* April 1st: Project submission
* April 5th: Project Demo

**Artifacts**

* Project Plan
* Deployment guidelines
* Final project report

**Acknowledgements**

- will be populated later